

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日

Date of Application:

2002年 8月29日

出願番号

Application Number:

特願2002-250281

[ST.10/C]:

[JP2002-250281]

出願人

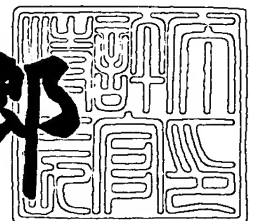
Applicant(s):

株式会社リコー

2003年 7月 3日

特許庁長官
Commissioner,
Japan Patent Office

太田信一郎



出証番号 出証特2003-3052848

【書類名】 特許願

【整理番号】 0203184

【提出日】 平成14年 8月29日

【あて先】 特許庁長官 太田 信一郎 殿

【国際特許分類】 G06F 17/30

【発明の名称】 単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体

【請求項の数】 20

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 真野 博子

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 伊東 秀夫

【特許出願人】

【識別番号】 000006747

【氏名又は名称】 株式会社リコー

【代表者】 桜井 正光

【代理人】

【識別番号】 100101177

【弁理士】

【氏名又は名称】 柏木 慎史

【電話番号】 03(5333)4133

【選任した代理人】

【識別番号】 100102130

【弁理士】

【氏名又は名称】 小山 尚人

【電話番号】 03(5333)4133

【選任した代理人】

【識別番号】 100072110

【弁理士】

【氏名又は名称】 柏木 明

【電話番号】 03(5333)4133

【手数料の表示】

【予納台帳番号】 063027

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9808802

【包括委任状番号】 0004335

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体

【特許請求の範囲】

【請求項 1】 文字列の入力を受付ける入力手段と、
この受付けた文字列から単語を抽出する単語抽出手段と、
この抽出した各単語について所定の文書群の各文書における特定の部位での出現の度合いを計算する出現度計算手段と、
を備えている単語出現度計算装置。

【請求項 2】 前記出現度計算手段は、前記特定の部位を前記文書における見出しとする、請求項 1 に記載の単語出現度計算装置。

【請求項 3】 前記出現度計算手段は、前記特定の部位を前記文書における要約とする、請求項 1 に記載の単語出現度計算装置。

【請求項 4】 前記出現度計算手段は、前記特定の部位を前記文書における見出し及び要約とする、請求項 1 に記載の単語出現度計算装置。

【請求項 5】 前記出現度計算手段は、“前記文書群で前記単語が前記見出しで出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項 2 に記載の単語出現度計算装置。

【請求項 6】 前記出現度計算手段は、“前記文書群で前記単語が前記要約で出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項 3 に記載の単語出現度計算装置。

【請求項 7】 前記出現度計算手段は、“前記文書群で前記単語が前記見出し又は前記要約で出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する、請求項 4 に記載の単語出現度計算装置。

【請求項 8】 前記出現度計算手段は、“（単語が見出しで出現する文書数 / 単語の出現する全文書数） + （単語が要約で出現する文書数 / 単語の出現する全文書数）”を計算することにより前記出現の度合いを計算する、請求項

4 に記載の単語出現度計算装置。

【請求項 9】 前記特定の部位の種類について選択を受付ける選択手段を備え、

前記出現度計算手段は、この選択された特定の部位における前記出現の度合いを計算する、請求項 1 ～ 8 の何れかの一に記載の単語出現度計算装置。

【請求項 10】 請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、
前記出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と
この選出した検索語に適合する文書を前記文書群から選出する文書選出手段と
を備えている文書検索装置。

【請求項 11】 請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、
前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、
を備えているキーワード抽出装置。

【請求項 12】 請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、
前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、
前記出現の度合いを利用して前記文書から文を抽出して要約文とする要約作成手段と、
を備えている文書要約装置。

【請求項 13】 請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、
前記出現の度合いを利用して前記文書から分類キーワードを抽出する分類キーワード抽出手段と、
この抽出結果を利用して前記文書群の文書を分類する分類手段と、

を備えている文書分類装置。

【請求項 1 4】 文字列の入力を受付ける入力手段と、
この受付けた文字列から単語を抽出する単語抽出手段と、
この抽出した各単語について所定の第 1 の文書群の各文書における出現の度合いを計算する出現度計算手段と、
この出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と

、
この選出した検索語に適合する文書を前記第 1 の文書群とは別の第 2 の文書群から選出する文書選出手段と、
を備えている文書検索装置。

【請求項 1 5】 請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項 1 6】 請求項 1 0 又は 1 5 に記載の文書検索装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項 1 7】 請求項 1 1 に記載のキーワード抽出装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項 1 8】 請求項 1 2 に記載の文書要約装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項 1 9】 請求項 1 3 に記載の文書分類装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラム。

【請求項 2 0】 請求項 1 5 ～ 1 9 の何れかの一に記載のプログラムを記憶している記憶媒体。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、文書から単語を抽出してその単語の所定の文書群における出現度合いを計算する、単語出現度計算装置、文書検索装置、キーワード抽出装置、文書要約装置、文書分類装置、プログラム及び記憶媒体に関する。

【 0 0 0 2 】

【従来の技術】

文書を多数集積している文書データベースからユーザーの必要とする文書を探し出すには、ユーザーがひとつあるいは数個程度の単語からなるキーワードを入力し、そのキーワードに適合する文書を選出する方法が一般的である。しかし、ユーザーの利用目的によっては、単語でなく、文を検索要求としたい場合もある。検索要求が短文 2 ～ 3 文程度であれば、検索要求から助詞などの不要語を取り除いて検索語とすれば、ユーザーのもとめる文書を十分な検索精度で探し出すことができる。たとえば、特開 2001-142897 号公報では、検索要求から複数単語の連続を抽出し検索する方法が提案されている。

【 0 0 0 3 】

【発明が解決しようとする課題】

しかしながら、もっと長い検索要求、例えば、文書全体があたえられたような場合には、この方法では、検索語が多くなりすぎ、検索に多大な時間がかかるだけでなく、ノイズの多い検索となり、検索精度が低下することが多い。

【 0 0 0 4 】

例えば、「昨年」「一昨年」などの副詞的名詞は、ほとんどの場合、検索に有用でないが、こういった単語は、取り除かれる不要語としてもれなく定義するのが難しいという不具合がある。

【 0 0 0 5 】

また、長い検索要求でも許容するようになると、短いキーワードによる入力では比較的問題にならなかった、文体や語彙や内容領域が検索におよぼす影響が大きくなり、特に、検索対象文書と大きく異なる文体や語彙や内容領域を持つ検索要求が入力された場合、例えば、新聞記事を検索要求として特許公報を検索対象文書とするような場合には、検索精度の低下が見られるという不具合がある。例を挙げると、「発売」などの単語は、新聞記事には多くみられても特許公報に出てくることは少ないが、検索では一般に検索対象文書の文書データベースでの出現文書数の少ない単語を重要とみなすので、「発売」は重要語とみなされることになってしまう。

【 0 0 0 6 】

本発明の目的は、長い文書が入力された場合でも、文書検索等に有用な重要語のみを選出できるようにすることである。

【 0 0 0 7 】

また、別の目的は、検索対象等となる文書群と大きく異なる文体や語彙や内容領域を持つ文章が入力された場合でも、適切な単語が選出されるようにすることである。

【 0 0 0 8 】

【課題を解決するための手段】

請求項 1 に記載の発明は、文字列の入力を受付ける入力手段と、この受付けた文字列から単語を抽出する単語抽出手段と、この抽出した各単語について所定の文書群の各文書における特定の部位での出現の度合いを計算する出現度計算手段と、を備えている単語出現度計算装置である。

【 0 0 0 9 】

したがって、文字列から抽出した単語の有用度を決定するのに、それぞれの単語が所定の文書群の中で出現する部位に着目し、その単語が出現する全出現数に比して所定の重要な部位に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 0 1 0 】

請求項 2 に記載の発明は、請求項 1 に記載の単語出現度計算装置において、前記出現度計算手段は、前記特定の部位を前記文書における見出しとする。

【 0 0 1 1 】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 0 1 2 】

請求項 3 に記載の発明は、請求項 1 に記載の単語出現度計算装置において、前

記出現度計算手段は、前記特定の部位を前記文書における要約とする。

【 0 0 1 3 】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 0 1 4 】

請求項 4 に記載の発明は、請求項 1 に記載の単語出現度計算装置において、前記出現度計算手段は、前記特定の部位を前記文書における見出し及び要約とする

。

【 0 0 1 5 】

したがって、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 0 1 6 】

請求項 5 に記載の発明は、請求項 2 に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記見出しで出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する。

【 0 0 1 7 】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを簡易、適切に計算することができる。

【 0 0 1 8 】

請求項 6 に記載の発明は、請求項 3 に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記要約で出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度

合いを計算する。

【 0 0 1 9 】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを簡易、適切に計算することができる。

【 0 0 2 0 】

請求項 7 に記載の発明は、請求項 4 に記載の単語出現度計算装置において、前記出現度計算手段は、“前記文書群で前記単語が前記見出し又は前記要約で出現する文書数 / 前記文書群で前記単語の出現する全文書数”を計算することにより前記出現の度合いを計算する。

【 0 0 2 1 】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【 0 0 2 2 】

請求項 8 に記載の発明は、請求項 4 に記載の単語出現度計算装置において、前記出現度計算手段は、“（単語が見出しで出現する文書数 / 単語の出現する全文書数） + （単語が要約で出現する文書数 / 単語の出現する全文書数）”を計算することにより前記出現の度合いを計算する。

【 0 0 2 3 】

したがって、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【 0 0 2 4 】

請求項 9 に記載の発明は、請求項 1 ～ 8 の何れかの一に記載の単語出現度計算装置において、前記特定の部位の種類について選択を受付ける選択手段を備え、前記出現度計算手段は、この選択された特定の部位における前記出現の度合いを計算する。

【 0 0 2 5 】

したがって、見出し、要約など複数種類の特定の部位から所望のものを選択して、その部位における単語の出現度合いを計算できるので、ユーザーの使い勝手を向上させることができる。

【 0 0 2 6 】

請求項 1 0 に記載の発明は、請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、この選出した検索語に適合する文書を前記文書群から選出する文書選出手段と、を備えている文書検索装置である。

【 0 0 2 7 】

したがって、文書検索の処理精度を向上させることができる。

【 0 0 2 8 】

請求項 1 1 に記載の発明は、請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、を備えているキーワード抽出装置である。

【 0 0 2 9 】

したがって、キーワード抽出の処理精度を向上させることができる。

【 0 0 3 0 】

請求項 1 2 に記載の発明は、請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書からキーワードを抽出するキーワード抽出手段と、前記出現の度合いを利用して前記文書から文を抽出して要約文とする要約作成手段と、を備えている文書要約装置である。

【 0 0 3 1 】

したがって、文書要約の処理精度を向上させることができる。

【 0 0 3 2 】

請求項 1 3 に記載の発明は、請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置と、前記出現の度合いを利用して前記文書から分類キーワードを抽出する分類キーワード抽出手段と、この抽出結果を利用して前記文書群の文書を分類する分類手段と、を備えている文書分類装置である。

【 0 0 3 3 】

したがって、文書分類の処理精度を向上させることができる。

【 0 0 3 4 】

請求項 1 4 に記載の発明は、文字列の入力を受付ける入力手段と、この受付けた文字列から単語を抽出する単語抽出手段と、この抽出した各単語について所定の第 1 の文書群の各文書における出現の度合いを計算する出現度計算手段と、この出現の度合いを利用して前記文書から検索語を選出する検索語選出手段と、この選出した検索語に適合する文書を前記第 1 の文書群とは別の第 2 の文書群から選出する文書選出手段と、を備えている文書検索装置である。

【 0 0 3 5 】

したがって、第 2 の文書群の文書と異なる文体や語彙や内容領域を持つ検索要求が入力された場合に、単語の有用度を決定するのに、入力した文書と同種の文体や語彙や内容領域を持つ文書からなる第 1 の文書群のそれぞれの単語の出現する度合いを計算すれば、入力文書から抽出した単語の第 1 の文書群で出現する度合いが第 2 の文書群で出現する度合いより大きいものは、有用度を下げることが可能となり、入力文書の同種文書に特有の単語を除くことができ、文書検索、キーワード抽出、文書要約、文書分類等の処理の精度が向上する。

【 0 0 3 6 】

請求項 1 5 に記載の発明は、請求項 1 ～ 9 の何れかの一に記載の単語出現度計算装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【 0 0 3 7 】

したがって、請求項 1 ～ 9 の何れかの一に記載の発明と同様の作用を奏する。

【 0 0 3 8 】

請求項 1 6 に記載の発明は、請求項 1 0 又は 1 5 に記載の文書検索装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【 0 0 3 9 】

したがって、請求項 1 0 又は 1 5 に記載の発明と同様の作用を奏する。

【 0 0 4 0 】

請求項 1 7 に記載の発明は、請求項 1 1 に記載のキーワード抽出装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【 0 0 4 1 】

したがって、請求項 1 1 に記載の発明と同様の作用を奏する。

【 0 0 4 2 】

請求項 1 8 に記載の発明は、請求項 1 2 に記載の文書要約装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【 0 0 4 3 】

したがって、請求項 1 2 に記載の発明と同様の作用を奏する。

【 0 0 4 4 】

請求項 1 9 に記載の発明は、請求項 1 3 に記載の文書分類装置の各手段が実行する処理をコンピュータに実行させるコンピュータに読取り可能なプログラムである。

【 0 0 4 5 】

したがって、請求項 1 3 に記載の発明と同様の作用を奏する。

【 0 0 4 6 】

請求項 2 0 に記載の発明は、請求項 1 5 ～ 1 9 の何れかの一に記載のプログラムを記憶している記憶媒体である。

【 0 0 4 7 】

したがって、その記憶しているプログラムにより請求項 1 5 ～ 1 9 の何れかの一に記載の発明と同様の作用を奏する。

【 0 0 4 8 】

【発明の実施の形態】

〔発明の実施の形態〕

本発明の一実施の形態を発明の実施の形態 1 として説明する。

【 0 0 4 9 】

図 1 は、本実施の形態である文書検索装置 1 の電氣的な接続を示すブロック図

である。図 1 に示すように、文書検索装置 1 は、P C などのコンピュータであり、各種演算を行ない文書検索装置 1 の各部を集中的に制御する C P U 2 と、各種の R O M や R A M からなるメモリ 3 とが、バス 4 で接続されている。

【 0 0 5 0 】

バス 4 には、所定のインターフェイスを介して、ハードディスクなどの磁気記憶装置 5 と、マウスやキーボードなどで構成される入力装置 6 と、L C D や C R T などの表示装置 7 と、光ディスクなどの記憶媒体 8 を読取る記憶媒体読取装置 9 とが接続され、また、インターネットなどのネットワーク 1 0 と通信を行なう所定の通信インターフェイス 1 1 が接続されている。なお、記憶媒体 8 としては、C D や D V D などの光ディスク、光磁気ディスク、フレキシブルディスクなどの各種方式のメディアを用いることができる。また、記憶媒体読取装置 9 は、具体的には記憶媒体 8 の種類に応じて光ディスクドライブ、光磁気ディスクドライブ、フレキシブルディスクドライブなどが用いられる。

【 0 0 5 1 】

磁気記憶装置 5 には、この発明のプログラムを実現する情報変換プログラムが記憶されている。この情報変換プログラムは、記憶媒体 8 から記憶媒体読取装置 9 により読取るか、あるいは、インターネットなどのネットワーク 1 0 からダウンロードするなどして、磁気記憶装置 5 にインストールしたものである。このインストールにより文書検索装置 1 は動作可能な状態となる。この文書検索プログラムは、特定のアプリケーションソフトの一部をなすものであってもよい。また、所定の O S 上で動作するものであってもよい。

【 0 0 5 2 】

図 2 に示すように、この文書検索装置 1 をサーバコンピュータ 1 4 として実施し、このサーバコンピュータ 1 4 と端末装置 1 2 とをネットワーク 1 3 を介して接続して、端末装置 1 2 からサーバコンピュータ 1 4 を操作できるようにしてもよい。この場合に、端末装置 1 2 は、パーソナルコンピュータ、携帯情報端末（P D A）、携帯電話などの情報処理装置として実施することができる。また、ネットワーク 1 3 は、無線、有線及び放送波のいずれを用いたものでもよく、例えば、L A N、W A N、インターネット、アナログ電話網、デジタル電話網（I S

DN)、PHS(パーソナルハンディホンシステム)網、携帯電話網、衛星通信網などを利用することができる。

【0053】

以下では、文書検索プログラムに基づいて文書検索装置1が行なう処理の内容について説明する。

【0054】

図3は、文書検索プログラムで実現される文書検索装置1の機能を説明する機能ブロック図である。文書検索装置1は、検索要求となる文章の入力を受付ける検索要求入力部21、検索語候補を抽出して、その検索語としての有用度を算出する検索語選出部22、検索語候補の指定部位出現度を計算する指定部位出現度計算部23、文書選出部24、文書出力部25、及び、文書データベース26等より構成される。文書データベース26は磁気記憶装置5に構築されるものであっても、文書検索装置1の外部に構築されるものであってもよい。

【0055】

図4は、文書検索プログラムに基づいて文書検索装置1が実行する処理のフローチャートである。まず、検索要求入力部21により、ユーザーがキーボード等で検索要求となる文章の文字列を入力する(ステップS1)。ステップS1により入力手段を実現する。この例では、「A社は、昨日、新しいプリンター AcmePrinter を発売した。」という新聞記事からの引用文を検索要求として入力したものとして説明する。

【0056】

かかる入力があると(ステップS1のY)、検索語選出部22は、入力された文章の文字列を所定の単語辞書により形態素解析して単語に分解する(ステップS2)。さらに、用意された不要語表に、この抽出した単語が登録されていれば不要語として削除して、残りの単語を検索語候補とする(ステップS3)。例えば、上の検索要求なら、「は」や「を」や「した」が不要語として削除され、「A社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補として残る。このステップS2、S3により単語抽出手段を実現している。

【0057】

さらに、検索語選出部 22 は、この各検索語候補について、検索語としての有用度を算出する。これには、例えば、以下の (1) 式を用いることができる。

【0058】

検索語の有用度 = 単語の重み …… (1)

ここで、「単語の重み」は、一般的には、
“ \log (全文書数 / 単語の出現文書数)” により求めることができる。すなわち、文書データベース 26 に登録されている文書群の中で出現文書数の少ない単語は、有用であるとみなす。

【0059】

しかし、この文書検索装置 1 では、指定部位出現度計算部 23 が、それぞれの単語が検索対象文書である文書データベース 26 の文書群の文書中で出現する部位 (文書中の「見出し」に出現するか、「要約」に出現するか、など) に着目し、その単語が指定の重要部位に出現する度合い (指定部位出現度) を単語の有用度に反映させる。

【0060】

例えば、文書の「見出し」を指定部位とした場合、指定部位出現度計算部 23 は、

指定部位出現度

= 単語が見出しで出現する文書数 / 単語の出現する全文書数
…… (2)

により、指定部位出現度を計算する。

【0061】

あるいは、文書の「要約」を指定部位とした場合には、

指定部位出現度

= 単語が要約で出現する文書数 / 単語の出現する全文書数
…… (3)

となる。

【0062】

あるいは、文書中の「見出し」及び「要約」の両方を指定部位とした場合は、

指定部位出現度

$$= \text{単語が見出し又は要約で出現する文書数} / \text{単語の出現する全文書数} \dots\dots (4)$$

としてもよい。

【0063】

さらに、上記(2)式と(3)式とを組み合わせ、

指定部位出現度

$$\begin{aligned} &= (\text{単語が見出しで出現する文書数} / \text{単語の出現する全文書数}) \\ &+ (\text{単語が要約で出現する文書数} / \text{単語の出現する全文書数}) \dots\dots (5) \end{aligned}$$

としてもよい。

【0064】

何れの手段でも、指定部位出現度を計算することにより、文書中における指定の重要部位で多く使われる単語を見分けることができる。その前提として、文書データベース26の電子化されている各文書について「見出し」「要約」などの各部分の範囲が文書中のどこからどこまでであるかを示すデータを持っているか、あるいは、各文書について「見出し」「要約」などの各部分ごとに各単語の出現数のデータを予め備えている必要がある。

【0065】

このようにして指定部位出現度計算部23が検索語候補の指定部位出現度を計算すると(ステップS4)、検索語選出部22は、指定部位出現度計算部23の算出した検索語候補の指定部位出現度を利用して検索語候補の有用度を計算して、検索語を抽出する(ステップS5)。ステップS4により出現度計算手段を、ステップS5により検索語選出手段を実現している。そして、ステップS1～ステップS4の機能により単語出現度計算装置を実現している。

【0066】

すなわち、(1)式から、

$$\text{検索語の有用度} = \text{単語の重み} \times \text{指定部位出現度} \dots\dots (6)$$

となる。

【0067】

あるいは、検索要求文章が長い場合には、

検索語の有用度

= 単語の重み × 指定部位出現度 × 検索要求文章内での出現回数

…… (7)

のように計算することもできる。

【0068】

このように、指定部位出現度を利用することにより、文書中の指定の重要部位で多く使われる単語を優先させることができる。

【0069】

この点につき、前述の文例で具体的に説明する。この文例は、「A社は、昨日、新しいプリンター AcmePrinter を発売した。」であり、「A社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補であった。

【0070】

下記の表1は、この各検索語候補である「単語」について、文書データベース26に登録されている文書群中で出現する文書の数「出現文書数」、その中でも文書の見出しで出現する文書の数「見出しでの出現文書数」、文書の要約で出現する文書の数「要約での出現文書数」として例示したものである。

【0071】

【表1】

単語	見出しでの出現文書数	要約での出現文書数	出現文書数
A社	22	22	30
昨日	0	10	16
新しい	2	8	24
AcmePrinter	8	8	12
発売	20	26	32

【0072】

この例において、(1)式で単語の有用度を計算すると、「昨日」は有用度が高いとみなされるが、(6)式で指定の重要部位に出現する度合いを利用して単

語の有用度を計算するなら、こういった単語の有用度は低く計算されることがわかる。

【 0 0 7 3 】

このように各検索語候補について有用度がもとまったら、ステップ S 5 において、検索語選出部 2 2 は、有用度の高い順に検索語候補を並べ、例えば、その上位 1 0 位を検索語として選出する。

【 0 0 7 4 】

そして、文書選出部 2 4 は、検索語選出部 2 2 が選出した検索語を用いて、文書データベース 2 6 を検索し、適合する文書を選定する（ステップ S 6）。ステップ S 6 により文書選出手段を実現している。

【 0 0 7 5 】

この選定された適合文書は、文書出力部 2 5 へ渡される。文書出力部 2 5 は、文書選出部 2 4 で選出した適合文書を、検索結果として出力する（ステップ S 7）。

【 0 0 7 6 】

また、部位種類指定部 2 7 は、指定部位出現度計算部 2 3 が前述のように指定部位出現度を計算する際の文書中の部位の種類（「見出し」か、「要約」か、あるいはその両方か）の選択を、ユーザーから受付ける。そして、この選択に応じて、指定部位出現度計算部 2 3 は（2）～（5）式の何れかにより指定部位出現度を計算する。

【 0 0 7 7 】

〔発明の実施の形態 2〕

別の実施の形態を発明の実施の形態 2 として説明する。

【 0 0 7 8 】

図 5 は、この実施の形態である文書検索装置 1 の機能ブロック図である。この文書検索装置 1 のハードウェア構成は、図 1、図 2 を参照して説明した発明の実施の形態 1 の場合と同様であり、詳細な説明は省略する。

【 0 0 7 9 】

この文書検索装置 1 が実施の形態 1 と相違するのは、文書群（第 1 の文書群）

を登録した第1の文書データベース31と、別の文書群（第2の文書群）を登録した第2の文書データベース32とを取り扱うこと、及び、指定部位出現度計算部23に代えてデータベース出現度計算部33を備えていることである。

【0080】

第1の文書データベース31、第2の文書データベース32は、磁気記憶装置5に構築されていても、文書検索装置1の外部に構築されていてもよい。第2の文書データベース32は前述の文書データベース26に相当するもので、検索対象文書からなる文書データベースである。第1の文書データベース31は、検索要求と同種の文体や語彙や内容領域を持つ文書からなる文書データベースである。この例では、第2の文書データベース32には特許公報の文書群がおさめられ、第1の文書データベース31には新聞記事の文書群がおさめられているものとする。

【0081】

図6は、文書検索プログラムに基づいて文書検索装置1が実行する処理のフローチャートである。

【0082】

ステップS11～S13の処理は、前述のステップS1～S3と同様である。ステップS11により入力手段を、ステップS12、S13により単語抽出手段を実現している。この例でも、検索要求入力部21により、「A社は、昨日、新しいプリンター AcmePrinter を発売した。」といった新聞記事からの引用文を入力したものとして説明する。ここでも、「A社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補として残る。そして、前述と同様に、各検索語候補について検索語としての有用度を（1）式により算出すると、第2の文書データベース32での出現文書数の少ない単語は、有用であるとみなされることとなる。

【0083】

しかし、本文書検索装置1では、データベース出現度計算部33が、それぞれの単語が、検索要求文書と同種の文体や語彙や内容領域を持つ文書からなる第1の文書データベース31で出現する頻度にも着目し、その頻度と、同じ単語が第

2の文書データベース32で出現する頻度との違いの度合い（データベース出現度）を、有用度に反映させる。そのために、まず、データベース出現度を計算する（ステップS14）。ステップS14により出現度計算手段を実現している。また、ステップS11～S14の機能により単語出現頻度計算装置を実現している。

【0084】

例えば、データベース出現度計算部33は、データベース出現度の算出のために、

データベース出現度

= 第2の文書データベースでの出現文書数 / 第2の文書データベース全文書数

- 第1の文書データベースでの出現文書数 / 第1の文書データベース全文書数

（ただし、値が負になる場合は、データベース出現度を0とする）

……（8）

のような計算をする。

【0085】

あるいは、

データベース出現度

= （第2の文書データベースでの出現文書数 / 第2の文書データベース全文書数） / （第1の文書データベースでの出現文書数 / 第1の文書データベース全文書数）

（ただし、値が1未満になる場合は、データベース出現度を1とする）

……（9）

のように計算してもよい。

【0086】

このようにして、第1の文書データベース31での単語の出現頻度と、第2の文書データベース32での単語の出現頻度とを用いてデータベース出現度を計算することにより、第2の文書データベース32では、比較的使われないが、第1

の文書データベース 3 1 ではよく使われる単語を選ばれにくくすることができる。

【 0 0 8 7 】

そして、検索語選出部 2 2 は、データベース出現度計算部 3 3 の算出するデータベース出現度を利用して単語の有用度を計算し、検索語を抽出する（ステップ S 1 5）。

【 0 0 8 8 】

すなわち、（1）式から、

検索語の有用度

$$= \text{単語の重み} \times \text{データベース出現度} \quad \dots\dots (10)$$

となる。

【 0 0 8 9 】

この点につき、前述の文例で具体的に説明する。この文例は、「A社は、昨日、新しいプリンター AcmePrinter を発売した。」であり、「A社」「昨日」「新しい」「プリンター」「AcmePrinter」「発売」が検索語候補であった。

【 0 0 9 0 】

下記の表 2 は、この各検索語候補である「単語」について、第 1 の文書データベース 3 1 に登録されている文書群中で出現する文書の数「第 1 の文書データベースでの出現文書数」、第 2 の文書データベース 3 2 に登録されている文書群中で出現する文書の数「第 2 の文書データベースでの出現文書数」として例示したものである。

【 0 0 9 1 】

【表 2】

単語	第1のデータベースでの 出現文書数	第2のデータベースでの 出現文書数
リコー	30	3
昨日	16	0
新しい	24	18
プリンター	12	10
AcmePrinter	6	0
発売	32	5

【0092】

この例において、例えば（1）式で単語の有用度を計算すると、「A社」や「発売」といった単語は有用度が高いとみなされるが、（10）式で単語の有用度を計算するなら、こういった単語の有用度は低く計算されることがわかる。

【0093】

ステップS15では、このように各検索語候補について有用度がもとまったら、検索語選出部22が、有用度の高い順に検索語候補をならべ、例えば、上位10位までを検索語として選出する。ステップS15により文書選出手段を実現している。

【0094】

ステップS16、S17の処理については、前述のステップS6、S7と同様であり、ここでは説明を省略する。

【0095】

なお、この例では、検索要求と検索対象とで文書の種類が異なる場合を例として説明した。すなわち、第1、第2の文書データベース31、32に登録されている文書群として新聞と特許公報とを例として挙げて説明した。この他に、同じ種類の文書であっても、検索要求と検索対象とで異なる分野に属する場合（例えば、特許公報であってもIPC分類が異なる場合など）や、検索要求と検索対象とが異なる著者の文書による場合などにも、この文書検索装置1は有益である。

【0096】

なお、実施の形態1と実施の形態2とを組み合わせる用いることもできる。す

なわち、単語の出現度をもとめるのに、指定部位出現度計算部 2 3 とデータベース出現度計算部 3 3 を併用するものである。

【0 0 9 7】

〔発明の実施の形態 3〕

別の実施の形態を発明の実施の形態 3 として説明する。

【0 0 9 8】

図 7 は、この実施の形態であるキーワード抽出装置 4 1 の機能ブロック図である。このキーワード抽出装置 4 1 のハードウェア構成は、図 1、図 2 を参照して説明した発明の実施の形態 1 の場合と同様であり、詳細な説明は省略する。

【0 0 9 9】

このキーワード抽出装置 4 1 では、図 1 のハードウェア構成で、記憶媒体 8 やネットワーク 1 0 からのダウンロードからインストールしたキーワード抽出プログラムが動作する。そして、キーワード抽出プログラムに基づく処理により、実施の形態 1 と同様な文書データベース 2 6 を扱い、実施の形態 1 と同様な機能を有する指定部位出現度計算部 2 3 と、キーワード抽出部 4 2 と、部位種類指定部 2 7 とを実現している。

【0 1 0 0】

図 8 は、キーワード検索プログラムに基づいてキーワード抽出装置 4 1 が実行する処理のフローチャートである。まず、キーワード抽出部 4 2 に、文書が入力されると（ステップ S 2 1 の Y）、その文書を対象に前述のステップ S 2、S 3 と同様の処理を行なう（ステップ S 2 2、S 2 3）。これにより、入力文書からキーワード候補となる単語が抽出される。ステップ S 2 1 により入力手段を、ステップ S 2、S 3 により単語抽出手段を実現している。

【0 1 0 1】

指定部位出現度計算部 2 3 は、各キーワード候補の指定部位出現度を、実施の形態 1 の場合と同様にして計算する（ステップ S 2 4）。ステップ S 2 4 により出現度計算手段を実現している。また、ステップ S 1 ～ S 4 により単語出現度計算装置を実施している。

【0 1 0 2】

そして、キーワード抽出部 4 2 は、指定部位出現度計算部 2 3 で算出された指定部位出現度を用いて単語の有用度を実施の形態 1 の場合と同様に求め、有用度の高い順にキーワード候補を並べて、例えば、上位 1 0 位までをキーワードとして選出する（ステップ S 2 5）。ステップ S 2 5 によりキーワード抽出手段を実現している。

【 0 1 0 3 】

このようにして、各文書の特徴をあらわすキーワードを的確に抽出することができる。

【 0 1 0 4 】

〔発明の実施の形態 4〕

別の実施の形態を発明の実施の形態 4 として説明する。

【 0 1 0 5 】

図 9 は、この実施の形態である文書要約装置 5 1 の機能ブロック図である。この文書要約装置 5 1 のハードウェア構成は、図 1、図 2 を参照して説明した発明の実施の形態 1 の場合と同様であり、詳細な説明は省略する。

【 0 1 0 6 】

このでは、図 1 のハードウェア構成で、記憶媒体 8 やネットワーク 1 0 からのダウンロードからインストールした文書要約プログラムが動作する。そして、文書要約プログラムに基づく処理により、実施の形態 3 と同様な文書データベース 2 6 を扱い、実施の形態 3 と同様な機能を有する指定部位出現度計算部 2 3 と、キーワード抽出部 4 2 とを実現している。実施の形態 3 と相違するのは、後述のような機能を備えた要約作成部 5 2 も実現している点である。

【 0 1 0 7 】

図 1 0 は、文書要約プログラムに基づいて文書要約装置 5 1 が実行する処理のフローチャートである。ステップ S 3 1 ～ S 3 4 は、前述のステップ S 2 1 ～ S 2 4 と同様の処理である。ステップ S 3 1 により入力手段を、ステップ S 3 2、S 3 3 により単語抽出手段を、ステップ S 3 4 により出現度計算手段を、それぞれ実現している。また、ステップ S 3 1 ～ S 3 4 の機能により単語出現度計算装置を実施している。そして、実施の形態 3 の場合と同様に、キーワード抽出部 4

2でキーワードを抽出する（ステップS35）。ステップS35によりキーワード抽出手段を実現している。

【0108】

このようにして、各文書の特徴をあらわすキーワードが得られるので、要約作成部52は、ステップS31で入力された文書から、このキーワードを所定程度多く含んでいる文だけを抽出し（ステップS36）、これらの文からなる文書を要約文として出力する（ステップS37）。例えば、キーワードを多く含む順に上位10位までの文を抽出することなどが考えられる。ステップS36により要約作成手段を実現している。

【0109】

このようにして、要約文を的確に作成することができる。

【0110】

〔発明の実施の形態5〕

別の実施の形態を発明の実施の形態5として説明する。

【0111】

図11は、この実施の形態である文書分類装置61の機能ブロック図である。この文書分類装置61のハードウェア構成は、図1、図2を参照して説明した発明の実施の形態1の場合と同様であり、詳細な説明は省略する。

【0112】

この文書分類装置61では、図1のハードウェア構成で、記憶媒体8やネットワーク10からのダウンロードからインストールした文書分類プログラムが動作する。そして、文書分類プログラムに基づく処理により、実施の形態1と同様な文書データベース26を扱い、実施の形態1と同様な機能を有する指定部位出現度計算部23、部位種類指定部27を実現している。さらに、後述のような機能を備えた分類キーワード選出部62と、分類部63も実現している。

【0113】

図12は、文書分類プログラムに基づいて文書分類装置61が実行する処理のフローチャートである。まず、分類キーワード選出部62に、文書が入力されると（ステップS41のY）、この文書を対象として前述のステップS2、S3と

同様の処理を実行する（ステップ S 4 2， S 4 3）。このようにして抽出された単語を分類キーワード候補とする。ステップ S 4 1 により入力手段を、ステップ S 4 2， S 4 3 により単語抽出手段を実現している。

【 0 1 1 4 】

次に、指定部位出現度計算部 2 3 は、各分類キーワード候補の指定部位出現度を計算する（ステップ S 4 4）。ステップ S 4 4 により出現度計算手段を実現している。また、ステップ S 4 1 ～ S 4 4 の機能により単語出現度計算装置を実施している。

【 0 1 1 5 】

そして、分類キーワード選出部 6 2 は、算出された指定部位出現度を用いて単語の有用度を実施の形態 1 の場合と同様に求め、有用度の高い順に分類キーワード候補を並べて、例えば、上位 1 0 位までを分類キーワードとして抽出する（ステップ S 4 5）。ステップ S 4 5 により分類キーワード抽出手段を実現している。

【 0 1 1 6 】

このようにして文書ごとに選出された分類キーワードに基づいて、分類部 6 3 は、文書を分類する（ステップ S 4 6）。ステップ S 4 6 により分類手段を実現している。これには、例えば、分類キーワードの単語ごとの有用度を要素とするベクトルを作成し、互いの内積を算出して、ベクトル間の距離を求め、距離の近いものどうしを同じ分類とすること等で実現する。これらについては周知の技術であるため、詳細な説明は省略する。このようにして分類された文書が得られる。

【 0 1 1 7 】

【発明の効果】

請求項 1 に記載の発明は、文字列から抽出した単語の有用度を決定するのに、それぞれの単語が所定の文書群の中で出現する部位に着目し、その単語が出現する全出現数に比して所定の重要な部位に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 1 1 8 】

請求項 2 に記載の発明は、請求項 1 に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 1 1 9 】

請求項 3 に記載の発明は、請求項 1 に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 1 2 0 】

請求項 4 に記載の発明は、請求項 1 に記載の発明において、文字列から抽出した単語の有用度を決定するのに、その単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを計算することで、文書検索、キーワード抽出、文書要約、文書分類等に有用な重要単語のみを特定することができ、これらの処理の精度が向上する。

【 0 1 2 1 】

請求項 5 に記載の発明は、請求項 2 に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出しに出現する度合いを簡易、適切に計算することができる。

【 0 1 2 2 】

請求項 6 に記載の発明は、請求項 3 に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である要約に出現する度合いを簡易、適切に計算することができる。

【 0 1 2 3 】

請求項 7 に記載の発明は、請求項 4 に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し

及び要約に出現する度合いを簡易、適切に計算することができる。

【 0 1 2 4 】

請求項 8 に記載の発明は、請求項 4 に記載の発明において、文字列から抽出した単語が所定の文書群の中で出現する全出現数に比して重要な部位である見出し及び要約に出現する度合いを簡易、適切に計算することができる。

【 0 1 2 5 】

請求項 9 に記載の発明は、請求項 1 ～ 8 の何れかの一に記載の発明において、見出し、要約など複数種類の特定の部位から所望のものを選択して、その部位における単語の出現度合いを計算できるので、ユーザーの使い勝手を向上させることができる。

【 0 1 2 6 】

請求項 1 0 に記載の発明は、文書検索の処理精度を向上させることができる。

【 0 1 2 7 】

請求項 1 1 に記載の発明は、キーワード抽出の処理精度を向上させることができる。

【 0 1 2 8 】

請求項 1 2 に記載の発明は、文書要約の処理精度を向上させることができる。

【 0 1 2 9 】

請求項 1 3 に記載の発明は、文書分類の処理精度を向上させることができる。

【 0 1 3 0 】

請求項 1 4 に記載の発明は、第 2 の文書群の文書と異なる文体や語彙や内容領域を持つ検索要求が入力された場合に、単語の有用度を決定するのに、入力した文書と同種の文体や語彙や内容領域を持つ文書からなる第 1 の文書群のそれぞれの単語の出現する度合いを計算すれば、入力文書から抽出した単語の第 1 の文書群で出現する度合いが第 2 の文書群で出現する度合いより大きいものは、有用度を下げることが可能となり、入力文書の同種文書に特有の単語を除くことができ、文書検索、キーワード抽出、文書要約、文書分類等の処理の精度が向上する。

【 0 1 3 1 】

請求項 1 5 ～ 1 9 の何れかの一に記載の発明は、請求項 1 ～ 1 4 の何れかの一

に記載の発明と同様の効果を奏する。

【 0 1 3 2 】

請求項 2 0 に記載の発明は、請求項 1 5 ～ 1 9 の何れかの一に記載の発明と同様の効果を奏する。

【図面の簡単な説明】

【図 1】

本発明の実施の形態 1 である文書検索装置の電気的な接続を示すブロック図である。

【図 2】

文書検索装置をサーバコンピュータとして端末装置と接続して使用する構成例のブロック図である。

【図 3】

文書検索装置の機能ブロック図である。

【図 4】

文書検索装置が行なう処理を説明するフローチャートである。

【図 5】

本発明の実施の形態 2 である文書検索装置の機能ブロック図である。

【図 6】

文書検索装置が行なう処理を説明するフローチャートである。

【図 7】

本発明の実施の形態 3 であるキーワード抽出装置の機能ブロック図である。

【図 8】

キーワード抽出装置が行なう処理を説明するフローチャートである。

【図 9】

本発明の実施の形態 4 である文書要約装置の機能ブロック図である。

【図 1 0】

文書要約装置が行なう処理を説明するフローチャートである。

【図 1 1】

本発明の実施の形態 5 である文書分類装置の機能ブロック図である。

【図 1 2】

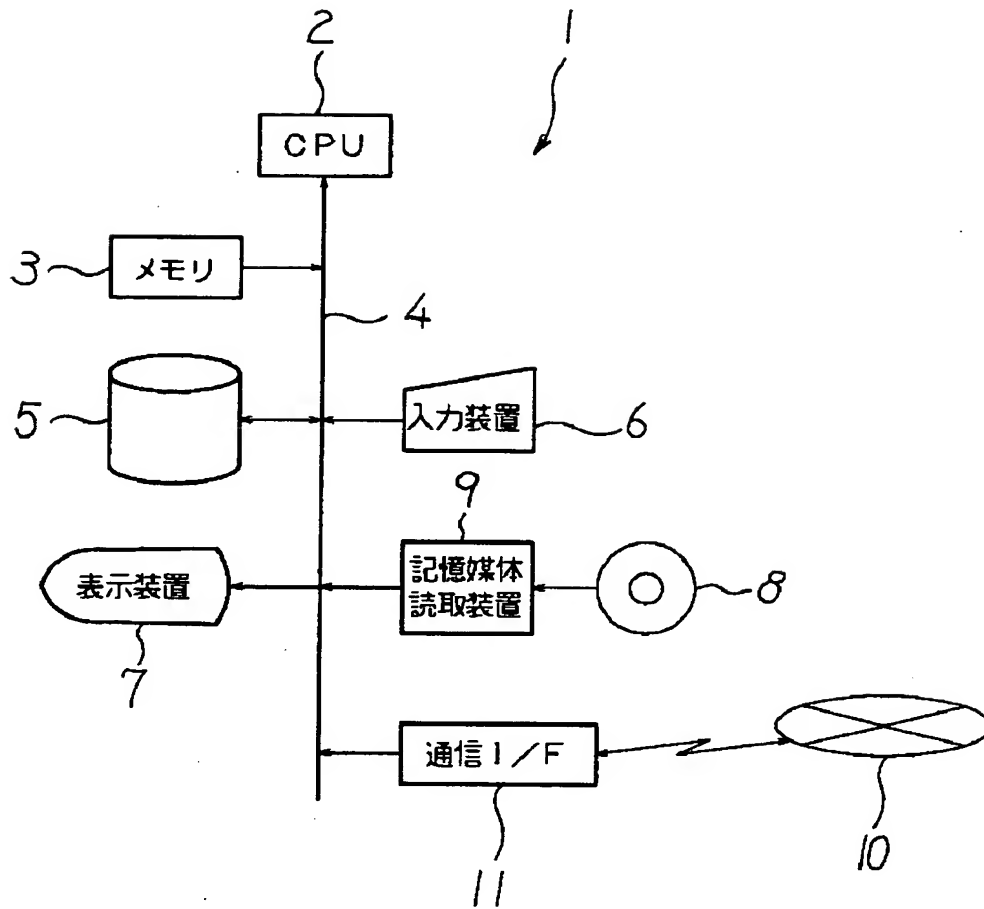
文書分類装置が行なう処理を説明するフローチャートである。

【符号の説明】

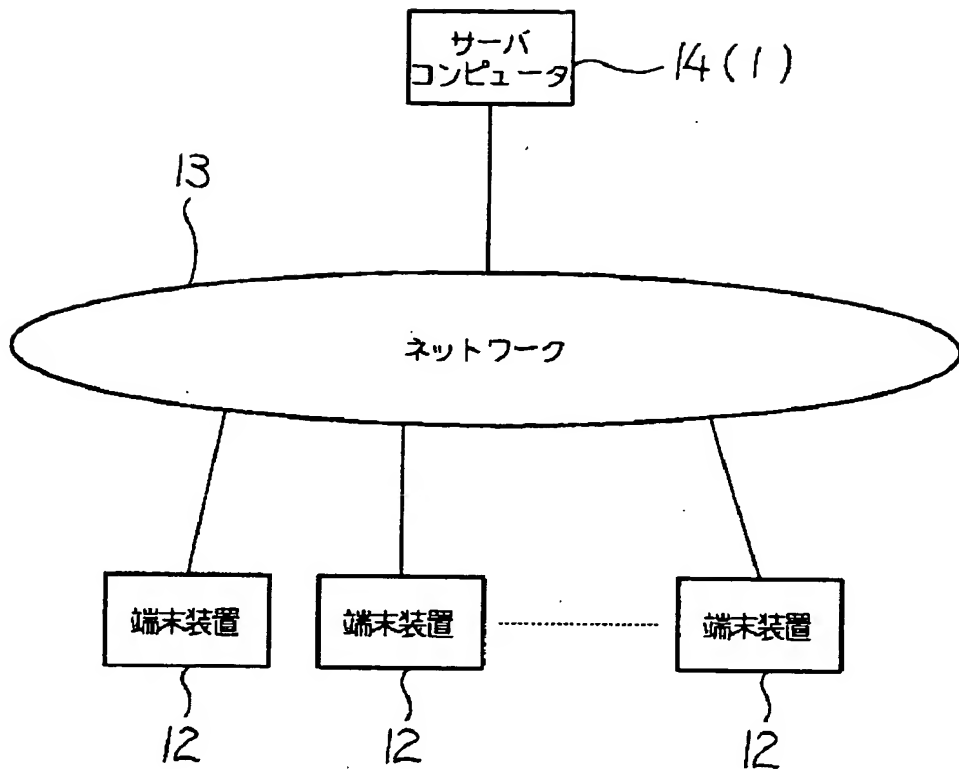
- 1 文書検索装置
- 8 プログラム
- 4 1 キーワード抽出装置
- 5 1 文書要約装置
- 6 1 文書分類装置

【書類名】 図面

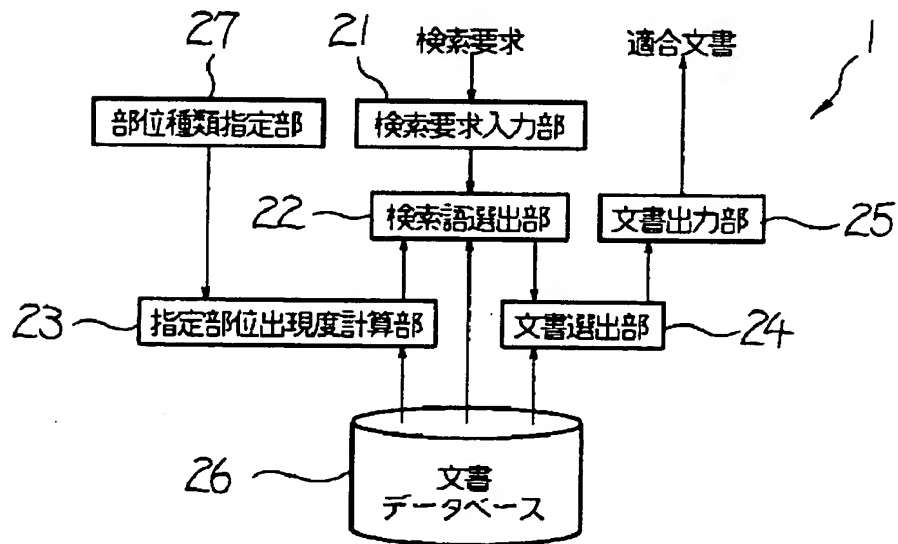
【図 1】



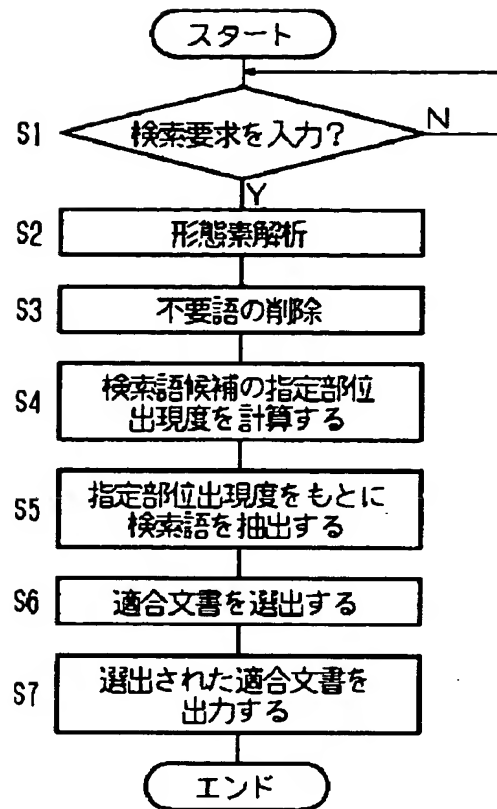
【図 2】



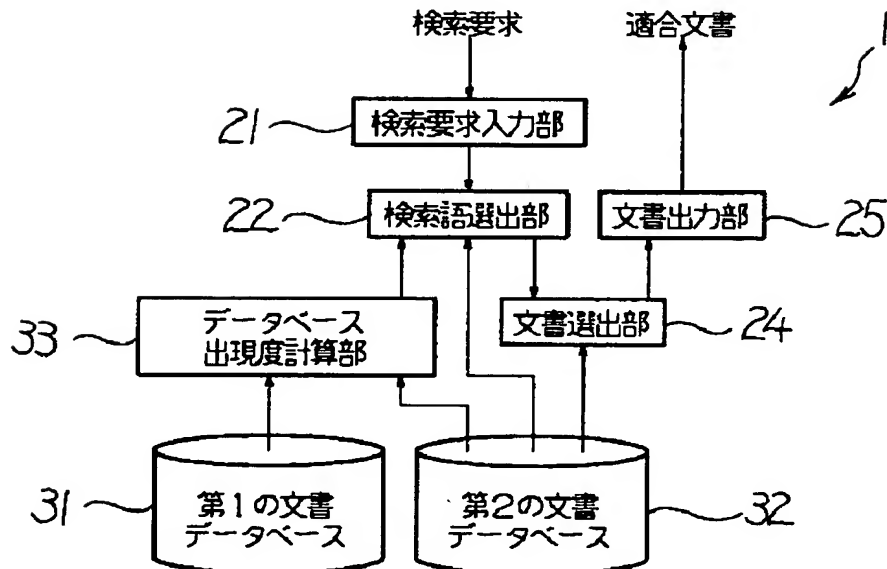
【図 3】



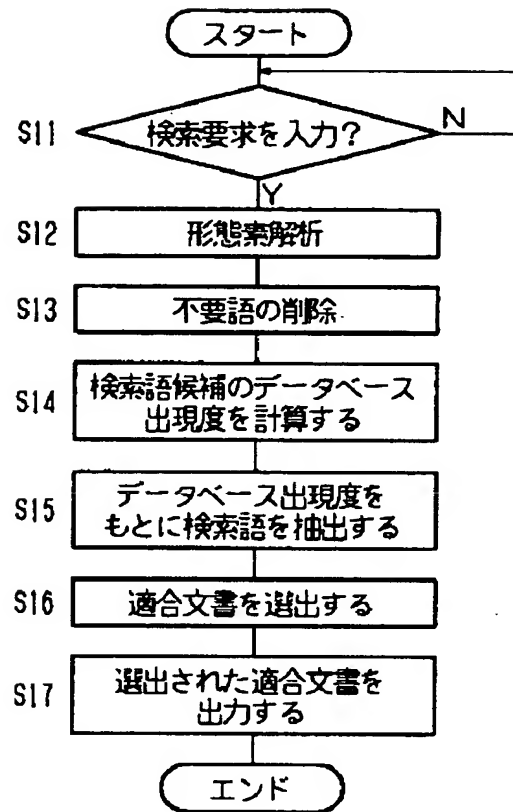
【図 4】



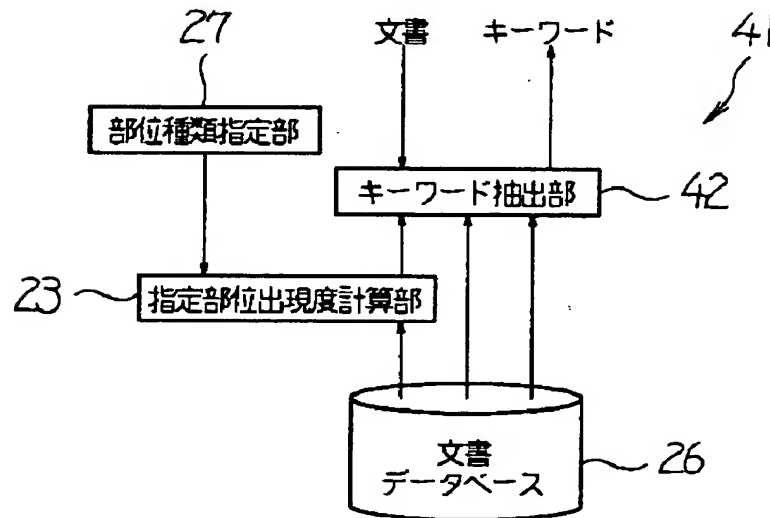
【図 5】



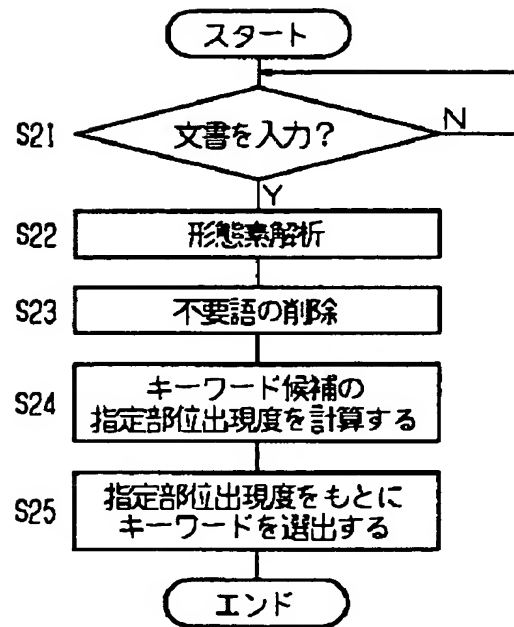
【図 6】



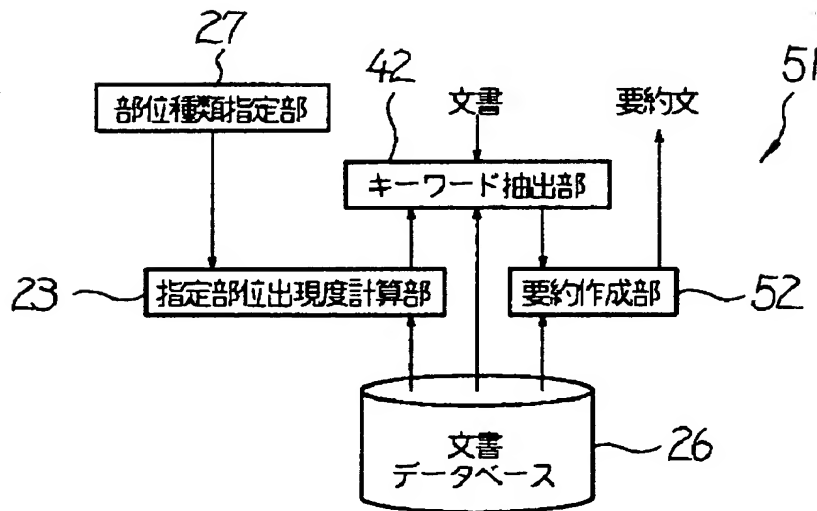
【図 7】



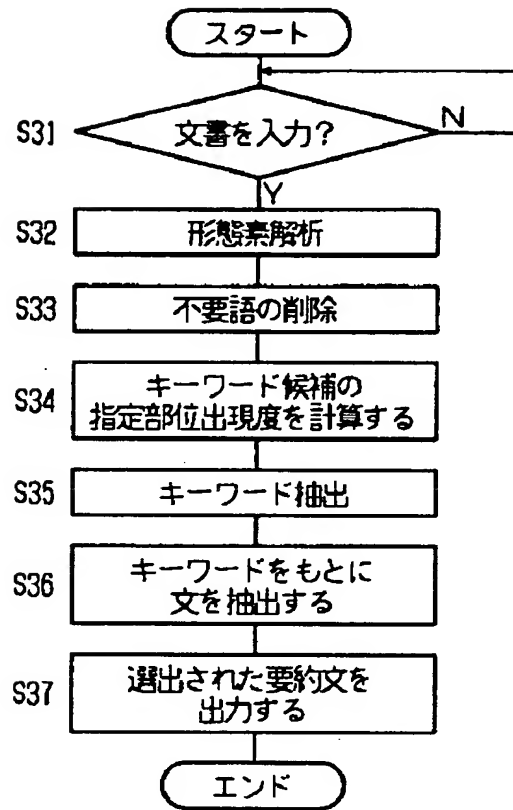
【図 8】



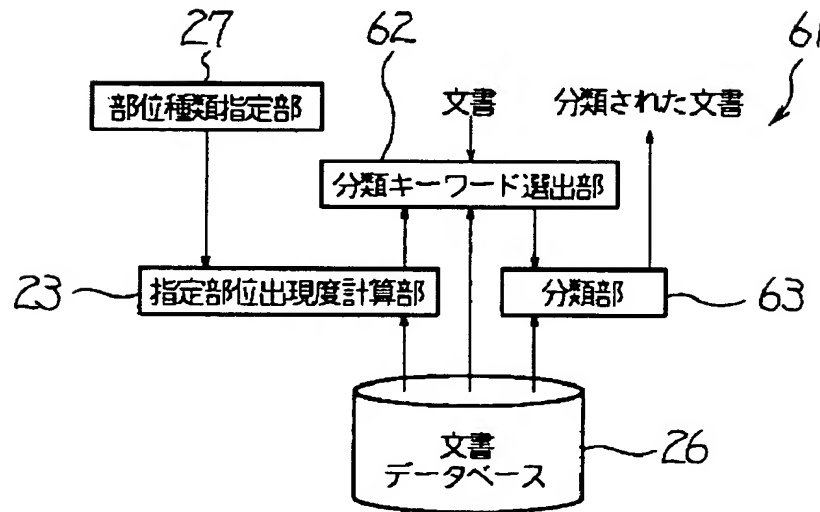
【図 9】



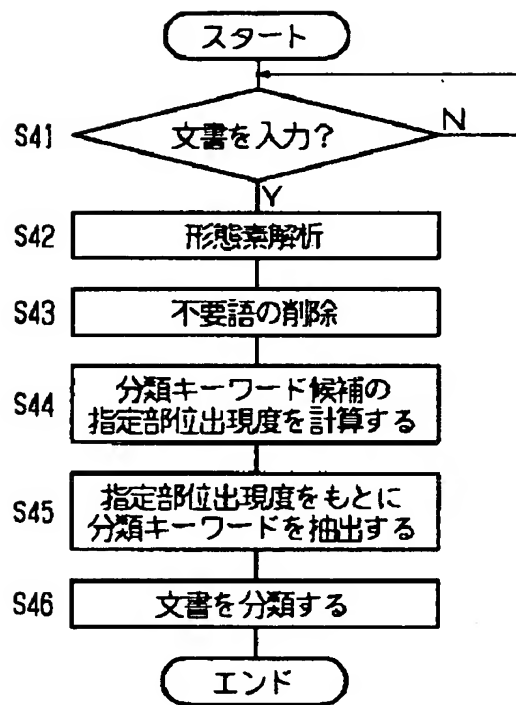
【図10】



【図11】



【図 12】



【書類名】 要約書

【要約】

【課題】 長い文書が入力された場合でも、文書検索に有用な重要語のみを選出できるようにして、文書検索の精度を向上する。

【解決手段】 検索要求の文書から不要語を除去した（ステップ S 2， S 3）後の各単語が、検索対象文書である文書中の見出し、要約などの重要部位に出現する度合い（指定部位出現度）を計算する（ステップ S 4）。そして、指定部位出現度をもとに検索語を抽出し（ステップ S 5）、この検索語から適合文書を選出する（ステップ S 6）。

【選択図】 図 4

出 願 人 履 歴 情 報

識別番号 [000006747]

1. 変更年月日 2002年 5月17日
[変更理由] 住所変更
住 所 東京都大田区中馬込1丁目3番6号
氏 名 株式会社リコー